Mining Least Association Rules of Degree Level Programs Selected by Students

Tutut Herawan[1], Zailani Abdullah[2], Wan Maseri Wan Mohd[1], A Noraziah[1]

[1]Faculty of Computer Systems and Software Engineering
Universiti Malaysia Pahang, Malaysia
Lebuhraya Tun Razak, 26300 Kuantan Pahang, Malaysia
[2]Department of Computer Science
Universiti Malaysia Terengganu
21030 Kuala Terengganu, Terengganu, Malaysia

{tutut,maseri,noraziah}@ump.edu.my, zailania@umt.edu.my

Abstract. One of the most popular and important studies in data mining is association rules mining. Generally, association rules can be divided into two categories called frequent and least. However, finding the least association rules is more complex and time consuming as compared to the frequent one. These rules are very useful in certain application domain such as determining the exceptional association between university's programs being selected by students. Therefore in this paper, we apply our novel measure called Definite Factors (DF) to determine the significant least association rules from undergraduate's program selection database. The dataset of computer science student for July 2008/2009 intake from Universiti Malaysia Terengganu was employed in the experiment. The result shows that our measurement can mine these rules and it is at par with the existing benchmarked Relative Support Apriori (RSA) measurement.
Keywords: Data Mining; Association rules; Significant Least, Measure; Educational Data.

1 Introduction

Data mining can be defined as the process of extracting hidden and useful information from large data repositories [1]. One of the emerging interdisciplinary research areas in data mining is educational data mining [2]. By definition, educational data mining is an application of suitable data mining techniques to analyze the educational data [3]. It aims at developing new methods that can discover the interesting information from educational settings, and used those methods to better understand the students, and their learning settings (http://www.educationaldatamining.org). The problem of association rules mining was first coined by [4] in an attempt for market-basket analysis. The classification of frequent or least items is based on the mechanism of support threshold. A set of items (itemset) is said to be frequent, if it appears more than minimum support count. The item (or itemset) support count is defined as a probability of item (or itemset) appears in the transaction. In addition, confidence is

another measure that always used together with support count. The confidence is defined as the probability of the rule's consequent (right side) that also contain the antecedent (left side) in the transaction. The association rule is said to be strong if it meets the minimum confidence threshold.

Least itemset is a set of item that is rarely occurred in the transactional database. It is also known as non-frequent, unusual, exceptional, abnormal, in-balance or sporadic itemset. In some applications domain, these itemsets are very important and in fact it can provide significant information such as air pollution level [5], relationship management [6], image processing [7], abnormal learning problems [8], educational data mining [9-11], text mining [12-13], information visualization [14-15], business process management [16] and many more. From the past literature, most of the tradition association rules mining algorithms [17-27] suffer in term of efficiency and evaluating the real datasets.

Educational data is one of the potential resources in discovering the significant least association rules. These rules can be very useful for higher authority personnel in assisting them to make right decision. For instance, in every July semester, our university receives approximately 160 students to enroll in computer science program. There are always the cases that the students are uncertain and taken for granted by combining with the various fields of interests. The research question is how to justify the student interests since there is no such field to be specified in the online application system. At the moment, if their choices are not selected, they will be offered to any program in the university according to programs availability.

Therefore, in this paper, we apply our novel measure called Definite Factors (DF) to detect the abnormal relationship among university's programs that have been selected by students. Indeed, DF will take into consideration the combination of both frequent and least university's program for generating the desired least association rules. We also employed our LP-Tree and LP-Growth algorithms [9] prior to produce the rules. In this study, the experiment was performed based on the 2008/2009 intake students' that have been offered in Bachelor of Information Technology (Software Engineering) at Universiti Malaysia Terengganu (UMT).

The rest of the paper is organized as follows. Section 2 describes the related work. Section 3 discusses the proposed method. This is followed by experiment tests in section 4. Finally, conclusion and future direction are reported in section 5.

2 Related Works

Nowadays, varieties of data mining method methods have been proposed in educational data mining. Romero et al. [28] suggested two categories of education data mining. The first category contains both statistics and visualization. The second one is web mining which can be divided into three parts. The first part covers clustering, classification, and outlier detection. The second part consists of association rule mining and sequential pattern mining. Finally, the third part is associated with text mining. It can be conclude that, the initial educational data mining is come into sight by analyzing the interaction between student and computer based on detailed logs of all their activities.

Baker et al. [2] proposed educational data mining into five different categories named prediction, clustering, relationship mining, distillation of data for human judge and discovery with model. Recently, discovery with models category is now become one of the most popular methods in educational data mining research. It deals with the sophisticated analysis such as discovering which learning materials sub-categories of students are the most beneficial [29], finding how different type of student behavior contribute to student's learning in different way [30] and revealing how variety of designing the intelligent tutor influence student's behavior over time [31].

Nowadays, only few attentions have been paid to extract least association rules from educational data. To the best of our knowledge, only one paper [8] is specifically discussed about least association rules. They applied the existing Rare Association Rules Mining (Apriori-like) algorithms to extract association rules from e-learning data. Their objective is to discover the information about infrequent student behavior. Four Apriori-based algorithms were employed to extract these rules named Apriori-Frequent [4], Apriori-Infrequent, Apriori-Inverse [19] and Apriori-Rare [32]. From the experiments, Apriori-Inverse and Apriori-Rare are proven more suitable in finding the least association rules.

In term of measurement least association rules, one of the popular measurement is Relative Support Apriori (RSA) proposed by [20]. RSA requires three (3) predefined measurements called 1st support, 2nd support and relative support (1st support > 2nd support). An item is said a least item if its support is less than 1st support and greater or more than 2nd support. A frequent item is an item having a support which equal or greater than 1st support. The least association rules are those rules that satisfied all the predefined supports. The main constrain of this algorithm is it increases the computational cost if the minimum relative support is set close to zero. In addition, determination of three predefined measurements is also another issue for this algorithm. Besides RSAA, the others approach to capture least association rules are Multiple Support Apriori [17], Matrix-based Scheme [18], Collective Support Apriori [33], etc.

3 Proposed Method

Throughout this section the set $I = \{i_1, i_2, \cdots, i_{|A|}\}$, for $|A| > 0$ refers to the set of literals called set of items, $W = \{w_1, w_2, \cdots, w_{|A|}\}$, refers to the set of literals called set of weights with a non-negative real numbers, and the set $D = \{t_1, t_2, \cdots, t_{|U|}\}$, for $|U| > 0$ refers to the data set of transactions, where each transaction $t \in D$ is a list of distinct items $t = \{i_1, i_2, \cdots, i_{|M|}\}$, $1 \le |M| \le |A|$ and each transaction can be identified by a distinct identifier TID.

3.1 Definition

In order to easily comprehend our measurement, some required definitions together with a sample transactional data are presented.

Definition 1. *A set* $X \subseteq I$ *is called an itemset. An itemset with k-items is called a k-itemset.*

Definition 2. *The support of an itemset* $X \subseteq I$, *denoted* $\text{supp}(X)$ *is defined as a number of transactions contain X.*

Definition 3. *Let* $X, Y \subseteq I$ *be itemset. An association rule between sets X and Y is an implication of the form* $X \Rightarrow Y$, *where* $X \cap Y = \phi$. *The sets X and Y are called antecedent and consequent, respectively.*

Definition 4. *The support for an association rule* $X \Rightarrow Y$, *denoted* $\text{supp}(X \Rightarrow Y)$, *is defined as a number of transactions in D contain* $X \cup Y$.

Definition 5. *The confidence for an association rule* $X \Rightarrow Y$, *denoted* $\text{conf}(X \Rightarrow Y)$ *is defined as a ratio of the numbers of transactions in D contain* $X \cup Y$ *to the number of transactions in D contain X. Thus*

$$\text{conf}(X \Rightarrow Y) = \frac{\text{supp}(X \Rightarrow Y)}{\text{supp}(X)}.$$

Definition 6. (Definite Factor). *Definite Factor is a formulation of exploiting the support difference between itemsets with the frequency of an itemset against a baseline frequency. The baseline frequency of itemset is presumed as statistically independence.*

The Definite Factor denoted as DF and

$$DF(I) = |P(X) - P(Y)| \times \frac{P(X \cup Y)}{P(X)P(Y)}$$

It also can be expressed as

$$DF(I) = |\text{supp}(X) - \text{supp}(Y)| \times \left(\frac{\text{supp}(X \Rightarrow Y)}{\text{supp}(X) \times \text{supp}(Y)} \right)$$

3.2 Construct Definite Least Association Rules

Rule is classified as Definite Least Association Rules (DLAR) if it fulfilled two conditions. First, DF of association rule must be greater than the predefined minimum DF. The range of min-DF is in between 0 and 1. Second, the antecedent and consequence of association rule must represent either Least Items or Frequent Items, respectively. The computation of DF of each association rule is employed from Definition 6. The complete procedure to construct the DLAR algorithm is as follows.

DLAR Algorithm
1: Specify $DF^{\min}$
2: for $(DI_a \in DefiniteItemset)$ do
3: for $(DFI_i \in DI_a \cap FrequentItems)$ do
4: for $(DLI_i \in DI_a \cap LeastItems)$ do
5: Compute $DF(DFI_i, DLI_i)$
6: if $(DF(DFI_i, DLI_i) > DF^{\min})$ do

```
7:                    Insert DLAR(DFI_i, DLI_i)
8:                end if
9:            end for loop
10:       end for loop
11: end for loop
```

Fig. 1: DLAR Algorithm

4 Experimental Results

In this section, we do experiment tests with DF measurements. The weight of all association rules were assigned according to this measurement. These experiments were conducted on Intel® Core™ 2 Quad CPU at 2.33GHz speed with 4GB main memory, running on Microsoft Windows Vista. All algorithms have been developed using C# as a programming language. We evaluate the proposed measurement to 2008/2009 intake students in computer science program. The data was obtained from Division of Academic, Universiti Malaysia Terengganu in a text file and Microsoft excel format. There were 160 students involved and their identities were removed due to the confidentiality agreement. In the original set of data, it consists of 35 attributes and the detail information were explained in 10 tables in Microsoft excel format.

Here, 8 chosen university programs by the students are extracted according to the fix location in the original flat file. The actual location for each programs are based on the fix column. There were in total of 822 bachelors programs offered in Malaysian public universities for July 2008/2009 students' intake. From this figure, 342 bachelor programs were selected by our 160 students and it can be generalized into 47 unique general fields. In addition, LP-Tree and LP-Growth algorithm with DF measure (Abdullah, et al, 2010) are employed in the experiment. The total of 4,177 association rules was successfully extracted. Table 1 depicts top 10 of association rules based on the 3% of minimum support. Table 2 illustrates the meaning of association rules based on the Table 1.

Table 1. Top 10 of association rules sorted by DF in descending order and 100% confidence

No	Association rules	Supp of Antc	Supp of Consq	Supp of Itemset	Jaccard	Corr	RSA	DF
1	25→9	4.38	90.63	4.38	0.048	22.86	1.00	0.95
2	25→34	6.25	90.63	6.25	0.07	16.00	1.00	0.93
3	25→8	12.50	90.63	12.50	0.14	8.00	1.00	0.86
4	25→43	9.38	90.63	8.75	0.10	10.67	0.93	0.84
5	28→40	3.75	68.75	3.12	0.05	26.67	0.83	0.79
6	25→41	8.13	90.63	6.88	0.07	12.31	0.85	0.77
7	25→38	15.00	90.63	13.75	0.15	6.67	0.92	0.76
8	25→31	5.63	90.63	4.38	0.05	17.78	0.78	0.73
9	28→34	6.25	68.75	5.00	0.07	16.00	0.80	0.73
10	25 28→34	6.25	60.23	5.00	0.08	16.00	0.80	0.72

Table 2. Explanation of top 10 of positive association rules

No	Association rules	Explanation
1	25→9	The student chose Forestry program also chose Banking program
2	25→34	The student chose Forestry program also chose Nursing program
3	25→8	The student chose Forestry program also chose Art Design
4	25→43	The student chose Forestry program also chose Radiotherapy program
5	28→40	The student chose IT program also chose Psychology
6	25→ 41	The student chose Forestry program also chose Pure Sciences
7	25→38	The student chose Forestry program also chose Physiotherapy
8	25→31	The student chose Forestry program also chose Management
9	28→34	The student chose IT program also chose Nursing
10	25 28→34	The student chose Forestry and IT program also chose Nursing

The link of interest between the antecedent and consequence for the first rule until fifth rule is quite strange due to the contradiction in the field of study among the respective programs. The sixth rule is very realistic since both programs have a similarity in term of basic requirements, link of interest and nature of study. For the sixth until tenth rules, it is very hard and confused to explains, since there is no link of interest between the programs. From here we can see that the students have mixed up with several interests during choosing their preferred university's programs. Moreover, most of them had chosen Forestry program. In summary, there are existed exceptional association rules in the university's program selection database. This information is very important to give an overall idea about the student interests and how to channel them to a more appropriate university's program.

5 Conclusion

Mining least association rules is very useful to help the organization in making a right decision. In educational context, identifying the suitable program for prospect students is very troublesome and usually ends up with programs availability. Therefore, this paper employed the Definite Factors measure to the students' enrolment data of computer science program (intake 2008/2009) at University Malaysia Terengganu. The result shows that the applied measure can discover the significant least association rules. From the generated rules, 32% of the students that have been offered in computer science program are not within their program interests. Thus, effective monitoring process and analysis of these students are very important in helping them to adapt and finally enjoy with the current program.

Acknowledgement. This research is supported by Fundamental Research Grant Scheme (FRGS) from Ministry of Higher Education of Malaysia Vote RDU 110104.

References

1. Tan, P-N., Steinbach, M., and Vipin, K.: Introduction to Data Mining. Addison-Wesley. (2006)
2. Baker, R., and Yacef, K.: The State of Educational Data Mining in 2009: A Review and Future Visions. Journal of Education Data Mining, 1 (1), 3–17 (2009)
3. Romero, C., Ventura, S., and García, E.: Data Mining in Course Management Systems: Moodle Case Study and Tutorial. Computers & Education, 51 (1), 368−384 (2008)
4. Agrawal, R., Imielinski, T., and Swami, A.: Database Mining: A Performance Perspective. IEEE Transaction on Knowledge and Data Engineering, 5 (6), 914–925 (1993)
5. Mustafa, M.D., Nabila, N.F., Evans, D.J., Saman, M.Y., and Mamat, A.: Association Rules on Significant Rare Data using Second Support. International Journal of Computer Mathematics, 83 (1), 69−80 (2006)
6. Au, W.H. & Chan, K.C.C.: Mining Fuzzy ARs in a Bank-Account Database. IEEE Transactions on Fuzzy Systems, 11 (2), 238−248 (2003)
7. Aggarwal, C.C. & Yu, P.S.: A New Framework for Item Set Generation. The Proceedings of the ACM PODS Symposium on Principles of Database Systems, 18−24 (1998)
8. Romero, C., Romero, J.R., Luna, J.M., and Ventura S.: Mining Rare Association Rules from e-Learning Data. In Proceeding of The Third International Conference of Education Data Mining, 171−180 (2010)
9. Abdullah, Z., Herawan, T., and Deris, M.M.: Mining Significant Least Association Rules using Fast SLP-Growth Algorithm. In T.H. Kim and H. Adeli (Eds.): AST/UCMA/ISA/ACN 2010, LNCS, Springer-Verlag, 6059, 324–336 (2010)
10. Abdullah, Z., Herawan, T., and Deris, M.M.: Scalable Model for Mining Critical Least Association Rules. In Rongbo Zhu et al. (Eds.): ICICA 2010, LNCS, Springer-Verlag, 6377, 509–516 (2010)
11. Herawan, T., Vitasari, P., and Abdullah, Z.: Mining Interesting Association Rules of Student Suffering Mathematics Anxiety. In J.M. Zain et al. (Eds.): ICSECS 2011, CCIS, Springer-Verlag, 188, II, 495–508 (2011)
12. Herawan, T., Yanto, I.T.R., and Deris, M.M.: Soft Set Approach for Maximal Association Rules Mining. In D. Ślęzak et al. (Eds.): DTA 2009, CCIS, Springer-Verlag, 64, 163–170 (2009).
13. Herawan, T., and Deris, M.M.: A Soft Set Approach for Association Rules Mining. Knowledge Based Systems, 24 (1), 186–195 (2011)
14. Herawan, T., Yanto, I.T.R., and Deris, M.M.: SMARViz: Soft Maximal Association Rules Visualization. In H. Badioze Zaman et al. (Eds.): IVIC 2009, LNCS, Springer-Verlag, 5857, 664–674 (2009)
15. Abdullah, Z., Herawan, T., and Deris, M.M.: Visualizing the Construction of Incremental Disorder Trie Itemset Data Structure (DOSTrieIT) for Frequent Pattern Tree (FP-Tree). In H.B. Zaman et al. (Eds.): IVIC 2011, LNCS, Springer-Verlag, 7066, 183–195 (2011)
16. Huang, Z., Lu, X., and Duan, H.: Mining association rules to support resource allocation in business process management. Expert Systems with Applications, 38 (8), 9483–9490 (2011)
17. Kiran, R.U., and Reddy, P. K.: An Improved Multiple Minimum Support Based Approach to Mine Rare Association Rules. In Proceeding of IEEE Symposium on Computational Intelligence and Data Mining, 340–47 (2009).

18. Zhou L. and Yau, S.: Assocation Rule and Quantative Association Rule Mining among Infrequent Items. In Proceeding of MDM2007 in conjunction with ACM SIGKDD'07, Article No. 9 (2007)
19. Koh, Y.S. and Rountree, N.: Finding Sporadic Rules using Apriori-Inverse. In: Ho, T.B., Cheung, D., Liu, H. (Eds): PAKDD 2005, LNCS, Springer-Verlag, 3518, 97−106 (2005).
20. Yun, H., Ha, D., Hwang, B., and Ryu, K.H.: Mining Association Rules on Significant Rare Data using Relative Support. The Journal of Systems and Software 67 (3), 181−19 (2003)
21. Liu, B., Hsu, W., and Ma, Y.: Mining Association Rules with Multiple Minimum Supports. In Proceeding of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'99), 337−341 (1999)
22. Wang, K., Hee, Y., and Han, J.: Pushing Support Constraints into Association Rules Mining. IEEE Transaction on Knowledge and Data Engineering, 15(3) 642−658 (2003)
23. Tao, F., Murtagh, F., and Farid, M.: Weighted Association Rule Mining using Weighted Support and Significant Framework. In Proceeding of ACM SIGKDD'03, 661−666 (2003)
24. Ding, J.: Efficient Association Rule Mining among Infrequent Items. Ph.D. Thesis, University of Illinois at Chicago (2005)
25. Abdullah, Z., Herawan, T., Noraziah, A., and Deris, M.M.: Mining Significant Association Rules from Educational Data using Critical Relative Support Approach. Procedia Social and Behavioral Sciences, 28, 97–101 (2011)
26. Abdullah, Z., Herawan, T., and Deris, M.M.: An Alternative Measure for Mining Weighted Least Association Rule and Its Framework. In J.M. Zain et al. (Eds.): ICSECS 2011, CCIS, Springer-Verlag, 188, II, 475–485 (2011)
27. Abdullah, Z., Herawan, T., Noraziah, A., and Deris, M.M.: Extracting Highly Positive Association Rules from Students' Enrollment Data. Procedia Social and Behavioral Sciences, 28, 107–111 (2011)
28. Romero, C. and Ventura, S.: Educational Data Mining: A Survey from 1995 to 2005. Expert Systems with Applications 33, 135–146 (2007)
29. Beck, J.E. and Mostow, J.: How who should practice: Using Learning Decomposition to Evaluate the Efficacy of Different Types of Practice for Different Types of Students. In Proceeding of the 9th International Conference on Intelligent Tutoring Systems, 353–362 (2008)
30. Cocea, M., Hershkovitz, A. and Baker, R.S.J.D.: The Impact of Off-task and Gaming Behaviors on Learning: Immediate or Aggregate. In Proceedings of the 14th International Conference on Artificial Intelligence in Education, 507−514 (2009)
31. Jeong, H. and Biswas, G.: Mining Student Behavior Models in Learning by Teaching Environments. In Proceeding of the 1st International Conference on Educational Data Mining, 127−136 (2008)
32. Szathmary, L., Napoli, A., and Valtchev, P.: Towards Rare Itemset Mining. In Proceeding of Internationall Conference on Tools with Artificial Intelligence 305−312 (2007)
33. Selvi, C.S.K., and Tamilarasi, A.: Mining Association Rules with Dynamic and Collective Support thresholds. Internationall Journal on Open Problems Computational Mathematics, 2(3), 427−438 (2009)